AFRL-IF-RS-TR-2006-190
Final Technical Report
May 2006

DATA MINING ALGORITHMS WITH PSEUDOKNOT FREE CODES

Anthony Macula

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2006-190 has been reviewed and is approved for publication

APPROVED: /s/

THOMAS RENZ
Project Engineer

FOR THE DIRECTOR: /s/

JAMES A. COLLINS, Deputy Chief
Advanced Computing Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
MAY 2006	Final	Jan 05 – Jan 06

4. TITLE AND SUBTITLE

DATA MINING ALGORITHMS WITH PSEUDOKNOT FREE CODES

5a. CONTRACT NUMBER
FA8750-05-C-0031

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61101E

6. AUTHOR(S)

Anthony J. Macula and Morgan Bishop

5d. PROJECT NUMBER
230T

5e. TASK NUMBER
00

5f. WORK UNIT NUMBER
01

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Anthony Macula
36 Westview Cr
Geneseo NY 14454

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/IFTC
525 Brooks Rd
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2006-190

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for Public Release; distribution unlimited. PA# 06-374

13. SUPPLEMENTARY NOTES
AFRL Project Engineer: Thomas Renz, IFTC, Thomas.Renz@rl.af.mil

14. ABSTRACT
Synthetic DNA is proposed for use as an information storage media and three-dimensional structural material in nanotechnology. The main requirement of these systems is large collections of oligonucleotides that will not crosshybridize. The process of designing them has come to be known as DNA word, or DNA code, design. In this research, the DNA code generating software, SysDCode was developed and refined to include pseudoknot secondary structure, simulate hybridization assays, and applied to DNA hybridization algorithms. SynDCode has the ability to create new DNA codes with high binding specificity, filter existing codes through verification, and extend codes where specific oligonucleotides are essential to overall system construction. SynDCode's robust yet efficient computational model allows for the exploration of unparalleled search space spawning superior codes with higher binding specificity.

15. SUBJECT TERMS
DNA Computing, Oligo Design, Synthetic DNA

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					Thomas Renz
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPONE NUMBER (*Include area code*)
U	U	U	UL	22	

Table of Contents

List of Figures

1. Summary

Synthetic DNA is proposed for use as an information storage media and three-dimensional structural material in nanotechnology. Such applications of DNA require collections of oligonucleotides that do not crosshybridize. Thus, there is a need to efficiently create large collections of non-crosshybridizing oligonucleotides. The process of designing them has come to be known as DNA word, or DNA code, design. In this research, the DNA code generating software, SynDCode, was developed and refined to include pseudoknot secondary structure, simulate hybridization assays, and applied to DNA hybridization algorithms.

2. Introduction

SynDCode provides the means to create collections of synthetic DNA strands (i.e., a DNA code) with controlled properties such as resistance to crosshybridization. The user has the ability to verify the properties of an existing DNA code, expand a given DNA code, or create an entirely new DNA code. The models built into SynDCode allow for the specification of thermodynamic properties of the generated DNA code and for collections of concatenated combinations of strands taken from the generated code. SynDCode also can be used to construct DNA codes that do not disrupt external and functional oligonucleotides, e.g., priming sites and it can use to construct codes that contain important motifs, e.g., restriction sites.

SynDCode captures the key aspect of the nearest neighbor thermodynamic model for hybridized DNA duplexes in a computationally efficient way. All pairwise strand computations have complexity $O(n^2)$ where n is the length of the strand. This is significant improvement, as other DNA code software tools based in nearest neighbor thermodynamics have complexity $O(n^3)$.

3. Methods, Assumptions, Procedures

3.1 Complemented DNA Codes

Single strands of DNA are, abstractly, (A, C, G, T) -quaternary sequences, with the four letters denoting the respective nucleic acids. In this paper, when we write DNA molecules without indicating the direction, it is assumed that the direction is $5' \to 3'$. To obtain the reverse complement of a strand of DNA, first reverse the order of the letters and then substitute each letter with its canonical Watson-Crick complement. For example, the reverse complement of AACGTG is CACGTT . If $x = AACGTG$, then we let $\overline{x}$ denote its reverse complement $CACGTT$. A complemented DNA code C is a collection of Watson-Crick pairs of complementary DNA sequences, i.e.,

$C = \{\{x_i, \overline{x_i}\}\}$. The greatest energy of duplex formation is obtained when two sequences are reverse complements of one another and the DNA duplex formed is a *Watson-Crick (WC) duplex*. However, there are many instances when the formation of non-WC duplexes is energetically favorable. In this paper, a non-WC duplex is referred to as a *crosshybridized (CH) duplex.* A basic goal of DNA code design is to be assured that a fixed temperature can be found that is well above the melting point of all CH and well below the melting point of all WC duplexes that can form from strands in the code. It is also desirable for all WC duplexes to have melting points in a narrow range. A complemented DNA code with this property is said to have high binding specificity. High binding specificity is akin to a high signal-to-noise ratio.

3.2 SynDCode

SynDCode provides the means to create collections of synthetic DNA strands with controlled properties such as resistance to crosshybridization. The user has the ability to verify the properties of an existing DNA code, expand a given DNA code or create an entirely new DNA code. The models built into SynDCode allow for the specification of thermodynamic properties of the generated DNA code and for collections of concatenated combinations of strands taken from the generated code. SynDCode can be used to construct DNA codes that do not adversely interact with functional oligonucleotides external to the code, e.g., priming sites, and it can construct codes that contain important motifs, e.g., restriction sites.

The mathematical models built into SynDCode come from [1] and allow for the specification of thermodynamic properties of the generated DNA code and for collections of concatenated combinations of strands taken from the generated code. All pairwise strand computations, including the computation of the bound on free energy of formation, $G(x : y)$, of the x:y duplex have complexity $O(n^2)$. This is significant, as other DNA code software tools, including SLSDesigner [2] use $G(x : y)$ approximation algorithms with complexity $O(n^3)$. These more computationally complex algorithms give more accurate pairwise $G(x : y)$ computations, but pairwise accuracy is not necessarily the most important consideration when the primary objective is maximizing the size of a code for global thresholds.

It should be noted that distance functions (like ours in SynDCode) which are based on the theory of t-gap block isomorphic subsequences are generalizations of the classical Levenshtein insertion-deletion distance and are very different from the *h-distance* frequently discussed in the DNA computing literature. The h-distance is a simple translation of the Hamming distance which, when applied to modeling potential secondary structures between two DNA strands of length n, takes into account at most $2n$ possible secondary structures between them. In stark contrast, the mathematical methods embedded in SynDCode give information about *all* potential secondary structures between two DNA strands and provides this information in a computational efficient way.

3.2.1 Weighted Stem Measure of a DNA Duplex

The notation from previous reports [1] and [8] are used throughout. A natural simplification for formulating binding specificity is to base it upon the maximum number of WC (inter-strand, non-covalent hydrogen) base pair bonds between complementary letter pairs which may be formed between two oppositely directed strands. Let $x:\overleftarrow{y}$ denote the CH formed between x and $\overline{y}$ when $\overline{y}$ is the WC complement of y. Then an upper bound on this maximum number of base pair bonds that can form in the $x:\overleftarrow{y}$ duplex is, $\psi_{\Omega}^{1}(x,y)$, the maximum length of a common subsequence to x and y. This doesn't mean that x and $\overline{y}$ will form d base pair bonds in a hybridization assay; it just says they could never form more than d base pair bonds. In [1], this measure was denoted by $\psi_{\Omega}^{1}(x,y)$ where Ω is the constant function 1.

If the binding specificity were solely dependent on the number of base pair bonds, then DNA codes constructed by using $\psi_{\Omega}^{1}(x,y)$ as the constraint could be used in hybridization assays with assured high binding specificity. However, the state of the art model of DNA duplex thermodynamics is the Nearest Neighbor Model (NN). In the NN model, thermodynamic (e.g., free energy) values are assigned to *loops* rather than base pairs. Consider two oppositely directed DNA strands

$$x = 5'x_1, x_2, ..., x_i, ..., x_n 3'$$
$$\overleftarrow{y} = 3'\overline{y_1}, \overline{y_2}, ..., \overline{y_j}, ..., \overline{y_n} 5'$$

where $\overline{y_j}$ denotes the complement to base y_j. A *secondary structure* of the DNA duplex $x:\overleftarrow{y}$ is a sequence of pairs of *complementary* bases $\left((x_{i_r}, \overline{y_{j_r}})\right)$ where (x_{i_r}) and $(\overline{y_{j_r}})$ are subsequences of x and $\overleftarrow{y}$ respectively. Clearly the duplex $x:\overleftarrow{y}$ can have many secondary structures. An important issue is to understand *which* secondary structure is the most energetically favorable. The duplex $x:\overleftarrow{y}$ can have a *t-stem* if and only if there are strings $\alpha, \beta \prec (n)$ with $\alpha = [i, i+t-1], \beta = [j, j+t-i]$ with $x_\alpha = y_\beta$ where $y = 5'y_1, y_2, ..., y_n 3'$. A *maximal t-stem* is one that is not properly contained in another larger t'-stem. Every maximal t-stem contains $\max(t-j+1, 0)$ j-stems. In [1] an efficient means of computing, $\psi_{\Omega}^{t}(x,y)$, the maximum weighted sum of the common t-stems that can occur taken over all possible secondary structures for the $x:\overleftarrow{y}$ duplex is given. Its method of computation is the basis of SynDCode.

Suppose $\Omega = F$ assigns each 2-stem its *Nearest Neighbor* thermodynamic free energy value as given in Figure 1. The $(i,j)^{th}$ entry of Table 1 is the value of $F(i,j)$. For example, $F(C,T) = 1.28$. This function is defined by setting $F(i,j)$ to free energies for

2-stems given in [10]. So - $F(C,T)$ denotes the free energy associated with the $\begin{matrix} 5'CT3' \\ 3'GA5' \end{matrix}$ 2-stem.

F	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AC	0	1.44	0	0	0	0.52	0	0	0	0	0	0	0	0	0	0
AG	0	0.13	1.28	0	0	0	0	0	0	0	0	0	0	0	0	0
AT	0	0	0	0.88	0	0	0	0	0	0	0	0	0	0	0	0
CA	0	0	0	0	1.45	0	0	0	0	0	0	0	0	0	0	0
CC	0	0	0	0	0	1.84	0	0	0	0	0	0	0	0	0	0
CG	0	0	0	0	0.47	0.11	2.17	0	0	0	0	0	0	0	0	0
CT	0	0	0	0	0.12	0.32	0	1.28	0	0	0	0	0	0	0	0
GA	0	0	0	0	0	0	0	0	1.3	0.25	0	0	0	0	0	0
GC	0	0.59	0	0	0	1.11	0	0	0	2.24	0	0.27	0	0.25	0	0
GG	0	0	0.32	0	0	0	0.11	0	0	1.11	1.84	0.52	0	0	0	0
GT	0	0	0	0	0	0	0	0.13	0	0.59	0	1.44	0	0	0	0
TA	0	0	0	0	0	0	0	0	0	0	0	0	0.58	0	0	0
TC	0	0	0	0	0	0	0	0	0	0	0	0	0	1.3	0	0
TG	0	0	0.12	0	0	0	0.47	0	0	0	0	0	0	0	1.45	0
TT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 1: Thermodynamic weight of 2-stems and mismatched stacked pairs. Mismatched, but stabilizing stacked pairs are in green, perfect 2-stems are in orange

Let $x:\overleftarrow{y}$ be a CH duplex and let $\Delta G(x:\overleftarrow{y})$ and $\Delta G(x:\overleftarrow{x})$ be the NN computation of the free energy of the $x:\overleftarrow{y}$ CH and $x:\overleftarrow{x}$ WC duplexes respectively. In [1], it is shown that $-\psi_F^2(x,y) \le \Delta G(x:\overleftarrow{y})$ and $-\psi_F^2(x,\overleftarrow{x}) = \Delta G(x:\overleftarrow{x})$. SynDCode generates blueprints for WC pairs in complemented DNA codes for which $\psi_\Omega^t(x,y)$ is the basic measure of compliance.

4. Results, Discussion

4.1 SynDCode Inputs

The inputs to SynDCode can be roughly divided into four categories: codeword generation parameters, code constraint parameters, code catenation option(s) and code extension. The code generation inputs are: "Length of DNA Codewords", "Initial Markov Probability Parameter", "Size of Additional Interval", "Size of Probe Interval", "Undesired Substrings" and "Guanine Only in Complement Strands". The code constraint inputs are "Stem Sizes to Check", "Corresponding Stem Thresholds", "Include Stabilizing Mismatch Stacked Pairs", "Maximum CH Duplex Free Energy Upper Bound" and "WC Duplex Free Energy Bounds." The code concatenation option is either turned

on or off. The code extension options allow the user to verify, extend or verify and extend existing code. See Figure 2.

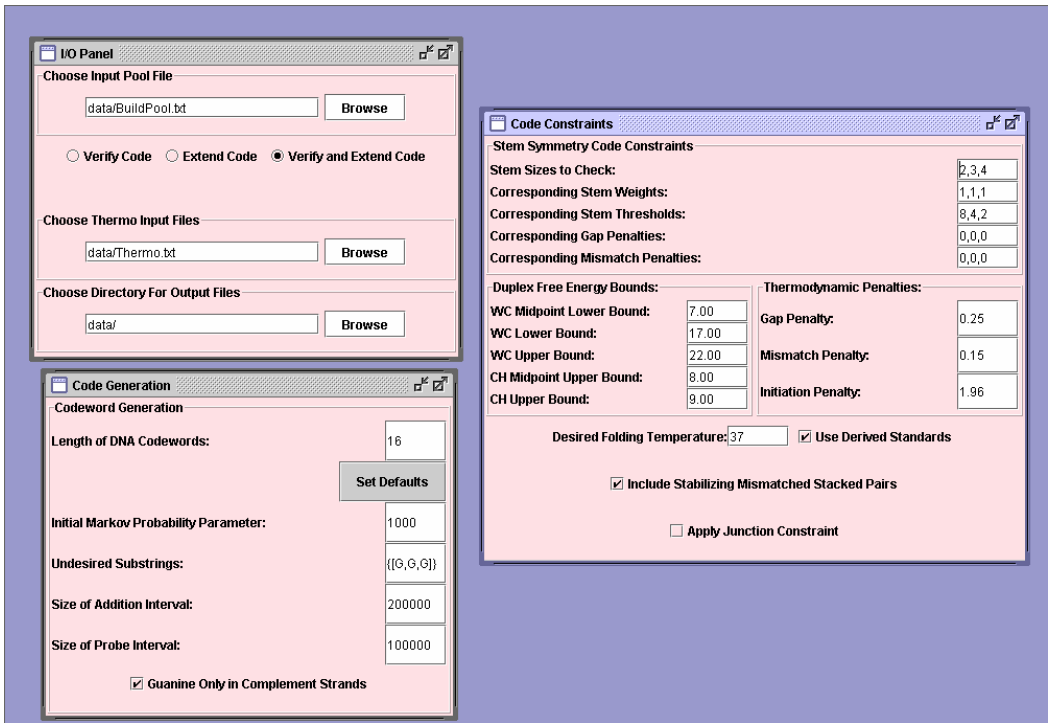

Figure 2: SynDCode GUI. Code Generation Windows

4.1.1 Codeword Generation Parameters

The input "Length of DNA Codewords" allows the user to select the desired length (in nucleotides) of each codeword and its complement. SynDCode generates a random code, but it doesn't (necessarily) do so by selecting from the uniform distribution. The input "Initial Markov Probability Parameter" allows the user to set the initial Markov probability parameter k to be used during candidate codeword generation. SynDCode generates candidate sequences $x_1, x_2, x_3, \ldots, x_n$ as reported earlier [8]. The parameters "Undesired Substrings", "Guanine Only in Complement Strands" are obvious and were also previously described in [8]. If a user selects the option "Guanine Only in Complement Strands" then x is generated in a manner similar to the above. That is, the value of x_1 is selected from $\{A, C, T\}$ with uniform probability. Then, the remaining entries are generated by a Markov chain given by a similar transition matrix, as in [8], where only $\{A, C, T\}$ are considered with parameter $k \geq 3$.

4.1.2 Code Constraint Parameters

The code constraint parameters are described in terms of $\psi_{\Omega}^{t}(x, y)$. Let $\psi^{t}(x, y)$ denote $\psi_{\Omega}^{t}(x, y)$ where $\Omega = 1$ and let $\psi_{\Omega}^{t}(x, y)$ be as described above. The input

"Stem Sizes to Check" allows the user to set the values of t for which $\psi^t(x,y)$ will be considered. The input *"Corresponding Stem Thresholds"* allow the user to set an upper bound s_i on the number of t_i -stems in a CH duplex. For example, if the user enters 2, 3, 4 (not required to be consecutive or in increasing order) in *"Stem Sizes"* and 8, 4, 2 in *"Corresponding Stem Threshold"* then every CH duplex $x : \overleftarrow{y}$ taken from the generated code will have $\psi^2(x,y) \le 8$, $\psi^3(x,y) \le 4$ and $\psi^4(x,y) \le 2$. The input "Maximum CH Duplex Free Energy Parameter" allows the user to set an upper bound for $\psi_F^2(x,y)$. Since $-\psi_F^2(x,y) \le \Delta G(x : \overleftarrow{y})$ this allows the user to bound the free energy of formation of all CH duplexes in the generated code. The "Maximum CH Midpoint Duplex Free Energy" parameter allows the user to bound the free energy of all CH duplexes between any strand and any half strand. In addition, all possible junctions $x_2 y_1$, where x_2 is the second half of strand x and y_1 is the first half of strand y , and any other strand z will have a free energy of formation $-\psi_F^2(x_2 y_1, z) \le 2 * \Delta G_{mid}(x : \overleftarrow{y})$. The inputs "WC Duplex Free Energy Bounds (Lower, Upper)" allows the user to ensure that all desired hybridized WC duplex have a free energy within a desired range. All WC pairs $x : \overleftarrow{x}$ in the generated code will have $\psi_F^2(x, \overleftarrow{x}) = -\Delta G(x, \overleftarrow{x})$ between the selected lower and upper values. For example, if 17.00 and 22.00 are selected as the lower and upper values then each WC pair $x : \overleftarrow{x}$ in the generated code with have $-22.00 \le \Delta G(x : \overleftarrow{x}) \le -17.00$. It is also known that single internal mismatches often times provide structural stability. See Figure 1. When the "Include Stabilizing Mismatch Stacked Pairs" option is selected, SynDCode includes these energetically favorable structures during computation of $\psi_F^t(x,y)$ and $\psi_F^2(x,y)$.

4.1.3 Pseudoknot Elimination

SynDCode restricts secondary structure by considering $\psi_F(x,x)$, and also predicts an upper bound on the optimal minimum free energy for an individual DNA sequence with any potential pseudoknot secondary structure $x_{\sigma_a \sigma_b} : x_{\sigma_{\overleftarrow{a'}} \sigma_{\overleftarrow{b'}}}$. In other words, every pseudoknot contains four distinct sections generally denoted as a,b,a',b'. See Figure 3.

Figure 3: Sequence Partition for Pseudoknot. Blue sections are a and a', red sections are b and b'. The sequence midpoint is marked by the yellow dot and the cut point marked by the green dot.

The algorithm operates by splitting the second half of the strand into two distinct partitions a' and b'. There are n/2 – 1 distinct partitions of the second half of a strand. Every partition is considered and each is referenced by the length of a' denoted with a cutpoint value. This pseudoknot free energy upper bound algorithm also has computational complexity $O(n^2)$. This model is sufficient for the purpose of DNA code design where exact structure prediction is not important. The main goal is to ensure that any pseudoknot formation is so unstable that it will not form. So, this function predicts the *absence* of pseudoknot secondary structure. There are n/2 – 1 distinct partitions of the second half of a strand. Every partition is considered and each is referenced by the length of a' denoted with a cutpoint value. The sections a and b are dynamically derived in order to produce an upper bounded free energy on the structure. In Figure 4, where the cutpoint is 3, a' is assigned to "CAC" and b' to "CTAAGTCGG." Then, the Watson-Crick complements of a' and b' are generated, concatenated together, and checked versus the unaltered first half of the strand producing an upper bound on the free energy of pseudoknot formation. The dot-parenthesis-bracket notation is generated above by applying a backtracking algorithm after computing the free energy. Any red open parenthesis represents the structural a section, any red closed parenthesis represents the a' section, any green closed bracket is part of the b section, and any open bracket is considered part of the b' section. Any dot represents an unbonded base. The matrix for the comparison where the cutpoint is 3 is in Figure 5.

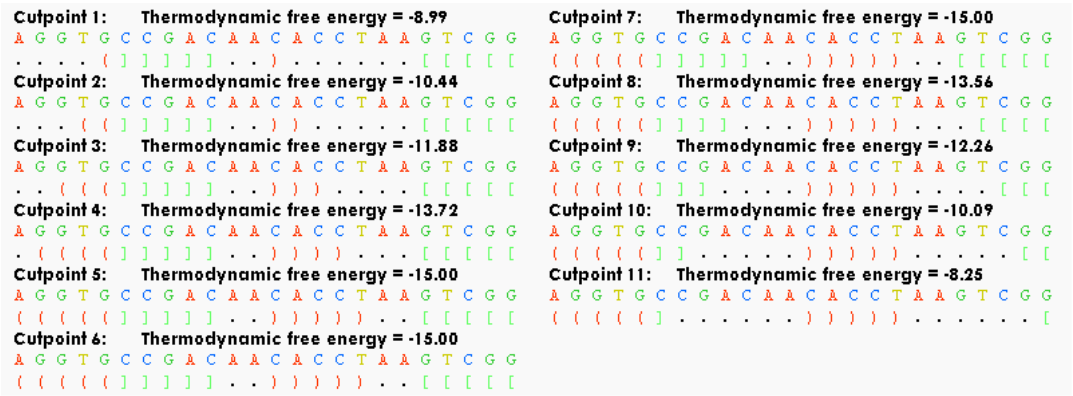

Figure 4: SynDCode Pseudoknot Alignment Output

			C	A	C	G	G	C	T	G	A	A	T	C
				CA	AC	CG	GG	GC	CT	TG	GA	AA	AT	TC
		0	0	0	0	0	0	0	0	0	0	0	0	0
A		0	0	0	0	0	0	0	0	0	0	0	0	0
G	AG	0	0	0	0	0	0	0	0	0	0	0	0	128
G	GG	0	0	0	0	0	0	0	0	0	0	0	0	128
T	GT	0	0	144	144	144	144	144	144	144	144	144	144	144
G	TG	0	0	144	289	289	289	289	289	289	289	289	289	289
C	GC	0	0	144	289	513	513	513	513	513	513	513	513	513
C	CC	0	0	144	289	513	697	697	697	697	697	697	697	697
G	CG	0	0	144	289	513	697	914	914	914	914	914	914	914
A	GA	0	0	144	289	513	697	914	1044	1044	1044	1044	1044	1044
C	AC	0	0	144	289	513	697	914	1044	1188	1188	1188	1188	1188
A	CA	0	0	144	289	513	697	914	1044	1188	1188	1188	1188	1188
A	AA	0	0	144	289	513	697	914	1044	1188	1188	1188	1188	***1188***

Figure 5: SynDCode Pseudoknot Alignment Matrix Output

An occurrence of most stable structure appears at cutpoint six where the free energy is -15 Kcal/Mol. The structure is given in Figure 6a.

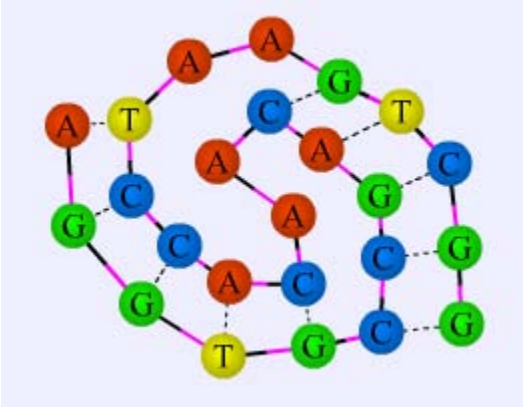

Figure 6a: Pseudoknot 2-d structure of that which was linearly represented as cutpoint 6 in Figure 5.

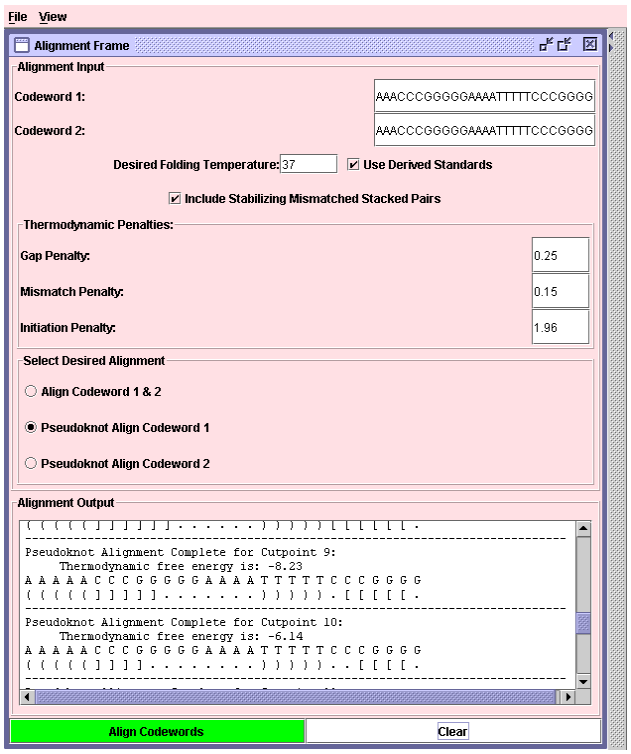

Figure 6b: SynDCode Aligner GUI

4.1.4 Example of Output and Junction Option

The strands in Figure 7 are part of code generated using the above parameters from Section 4.1.2 when the "Apply Junction Constraint" is selected. The WC pairs are labeled S_i and C_i .

Encoding	Reading/Probe
S_1= CCTTTTTTTTTTTTCC	C_1= GGAAAAAAAAAAAAGG
S_2= CCAATTTTCAAAACTT	C_2= AAGTTTTGAAAATTGG
S_3= TTAACCTTCCTTAAAA	C_3= TTTTAAGGAAGGTTAA
S_4=AATCCATTTTTACCAA	C_4=TTGGTAAAAATGGATT
S_5=CAACCAAAAACAAATT	C_5=AATTTGTTTTTGGTTG
S_6=TCCTCTTCCATCCTCC	C_6= GGAGGATGGAAGAGGA
S_7= ACCTCCACTTATTTAA	C_7= TTAAATAAGTGGAGGT
S_8=TACTCACACTTTACAA	C_8=TTGTAAAGTGTGAGTA
S_9= CTCAATTCTTTCTTTT	C_9= AAAAGAAAGAATTGAG
S_{10}=TAAACACCTATTCATT	C_{10}= AATGAATAGGTGTTTA
S_{11}=CTCTAAATCTAAACCT	C_{11}=AGGTTTAGATTTAGAG
S_{12}=ATCACAACATATTTCC	C_{12}=GGAAATATGTTGTGAT

Figure 7: A Coupled DNA Code

Here are the facts about this code:

1. Each of the 12 WC pairs have a free energy of formation parameter of between -22 and -17 KCAL per mole (i.e., $-22 \le -\psi_F^2(x, \overleftrightarrow{x}) \le -17$.)
2. No strand in the code contains the substring GGG or its complement CCC.
3. In each WC pair only one strand contains a G.
4. Each of the $\binom{24}{2} - 12$ potential CH pairs $x : \overleftarrow{y}$ with $x \neq y$ have a free energy of formation parameter that is not below -9 KCAL per mole (i.e.,. $-9 \le -\psi_?^2(x, y)$). Also each of the 24 CH pairs $x : \overleftarrow{x}$ have a free energy of formation parameter that is not below -9 KCAL per mole.
5. None of the $\binom{24}{2} - 12$ potential CH pairs have more than eight 2 -stems in any secondary structure (i.e.,. $\psi^2(x, y) \le 8.$)
6. None of the $\binom{24}{2} - 12$ potential CH pairs have more than four 3 -stems in any secondary structure (i.e.,. $\psi^3(x, y) \le 4.$)
7. None of the $\binom{24}{2} - 12$ potential CH pairs have more than two 4 -stems in any possible secondary structure (i.e.,. $\psi^4(x, y) \le 2.$)

When the "Apply Junction Constraint" option is selected the above complemented code satisfies additional constraints which we now describe. Using the encoding strands S_i with $1 \le i \le 10$ the following "combinatorial set" of 64 concatenated strands $X_1X_1X_3X_4X_5$ where $X_i = S_{2i}$ or S_{2i-1} can be constructed. The combinatorial set denotes all bit strings of length 6 if we let each S_{2i-1} denote "false" and each S_{2i} denote "true." To reduce crosshybridization potential between these longer concatenated strands and the probe strands C_i , our code satisfies further "catenation constraints." These catenation constraints largely follow from the "Maximum CH Midpoint Duplex Free Energy" function. To aid in the discussion of this constraint, consider the following set of *junction sequences* depicted in Figure 8.

Encoding	Junction	Encoding	Junction
s_1= CCTTTTTT-TTTTTTCC	S_1S_3 S_1S_4	s_6=TCCTCTTC-CATCCTCC	S_6S_7 S_6S_8
s_2= CCAATTTT-CAAAACTT	S_2S_3 S_2S_4	s_7= ACCTCCA-CTTATTTAA	S_7S_9 S_7S_{10}
s_3= TTAACCTT-CCTTAAAA	S_3S_5 S_3S_6	s_8=TACTCACA-CTTTACAA	S_8S_9 S_8S_{10}
s_4=AATCCATT-TTTACCAA	S_4S_5 S_4S_6	s_9= CTCAATTC-TTTCTTTT	S_9S_{11} S_9S_{12}
s_5=CAACCAAA-AACAAATT	S_5S_7 S_5S_8	s_{10}=TAAACACC-TATTCATT	$S_{10}S_{11}$ $S_{10}S_{12}$

Figure 8: Junction Sequence of a Coupled DNA Code

The junction sequence $S_i S_j$ where $j = i+1$, $i+2$ if i is even and $j = i+2, i+3$ if i is odd is the second half of sequence S_i and the first half of sequence S_j . For example, $S_1 S_4$ =TTTTTTCC AATCCATT. For the above code of 12 pairs of strands, we have a total of 20 such junction strands. If we add these junctions strands, we have a total of 44 strands under consideration. With the catenation constraint selected, SynDCode generates the 12 pairs of strands in Figure 7, so that not only are the conditions 1 to 7 above satisfied for the strands in Figure 7, but each of the possible $\binom{44}{2} - 12$ CH duplexes that can be formed from the strands in Table 3 also satisfy the same CH constraints. Thus the strands in Figure 7 were generated by SynDCode in such a way so that the strands in Figure 7 satisfy conditions 1 to 7 and the strands in Figure 8 satisfy:

8. Each of the $\binom{44}{2} - 12$ potential CH pairs have a free energy of formation that is not below -9 KCAL per mole (i.e.,. $-9 \le -\psi_F^2(x, y)$)

9. None of the $\binom{44}{2} - 12$ potential CH pairs have more than eight 2 -stems in any secondary structure (i.e.,. $\psi^2(x, y) \le 8.$)

10. None of the $\binom{44}{2} - 12$ potential CH pairs have more than four 3 -stems in any secondary structure (i.e.,. $\psi^3(x, y) \le 4.$)

11. None of the $\binom{44}{2} - 12$ potential CH pairs have more than two 4 -stems in any possible secondary structure (i.e.,. $\psi^4(x, y) \le 2.$)

5. Conclusions

5.1 Self-Assembly

The basic goal of DNA self-assembly is the autonomous formation of WC duplexes. The main issue is that when many strands are present in a solution crosshybridization can occur; leading to errors during the assembly process. SynDCode can be used to mitigate the CH problem and also to address the probabilistic and dynamic aspects of hundreds of strands in solution.

5.1.1 The Partition Function

Suppose we have a collection of strands $\{X_1, X_2, ..., X_{2n}\}$ where X_i, X_{i+1} are WC complements. For a given X_i we consider $2n-1$ possible duplexes that contain X_i as the possible states that X_i can be in. Let ΔG_{ij} be the free energy of formation of the duplex $X_i X_j$. In this application, we think of ΔG_{ij} only as a function of temperature. The ΔG_{ij} of $X_i X_j$ is computed by the latest version of SynDCode which allows the user to select the temperature of the assay. The thermodynamic values of stacked pairs at a given temperature T are computed by using the Gibbs formula

$$\Delta G = \Delta H - T\Delta S$$

where the values of ΔH and ΔS in Figure 9 come from [10].

Propagation sequence	ΔH° (kcal mol^{-1})	ΔS° (e.u.)	ΔG°_{37} (kcal mol^{-1})
AA/TT	−7.6	−21.3	−1.00
AT/TA	−7.2	−20.4	−0.88
TA/AT	−7.2	−21.3	−0.58
CA/GT	−8.5	−22.7	−1.45
GT/CA	−8.4	−22.4	−1.44
CT/GA	−7.8	−21.0	−1.28
GA/CT	−8.2	−22.2	−1.30
CG/GC	−10.6	−27.2	−2.17
GC/CG	−9.8	−24.4	−2.24
GG/CC	−8.0	−19.9	−1.84

Figure 9: Enthalpy and Entropy of Stacked Pairs at 37C

The probability that X_i is paired with X_{j_0} is given by the partition function

$$\frac{\exp(\Delta G_{ij_0}/RT)}{\exp(\Delta G_{ij_0}/RT) + \sum_{j \neq j_0} \exp(\Delta G_{ij}/RT)}.$$

To model self-assembly, the probability that in an ensemble of strands, *each* strand is paired with its complement and how long it takes to get to that state is important. This computational and simulation software is now part of SynDCode. SynDCode can be

used to select the optimal temperature at which the self assembly assay should occur. For example, consider a DNA Code with 2n strands. If the WC free energy $-\Delta G_{WC} \geq w$ and the CH free energy $-\Delta G_{CH} \leq c$, then given that each object has *access* to each state, the probability that each strand will find its complement in an ensemble is at least $\left(\frac{\exp(w/RT)}{\exp(w/RT)+(n-1)\exp(c/RT)}\right)^n$. Applying this analysis to the code in Figure 7 (first 10 pairs) which has $-\Delta G_{WC} \geq 17$ and $-\Delta G_{CH} \leq 9$ (at $310^\circ K$), then the probability that each strand pairs with its complement in the total ensemble is over 99.99%. The self-assembly hybridization of this code was simulated using SynDCode. See Figure 10.

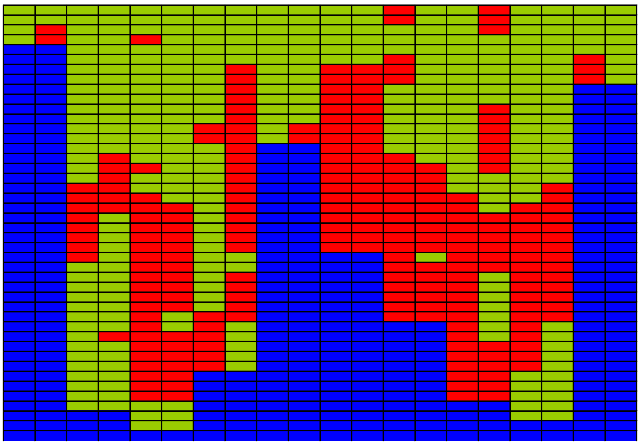

Figure 10: Epochal Self-Assembly Simulation. Each encoding *S* strand is accompanied with its complement probe strand *C* in its right-hand-side adjacent column. At the end of each epoch (row), every strand in the pool is either free (colored green), cross-hybridized bonded (red), or Watson-Crick bonded (blue).

5.1.2 Self-Assembly Simulation

Figure 10 illustrates the hybridization of 10 distinct DNA Watson-Crick pairs in silico with each column representing 1 strand and each row representing 1 epoch. Each encoding *S* strand is accompanied with its complement probe strand *C* in its right-hand-side adjacent column. For example, *S4* is addressed to column 7 with *C4* addressed to column 8. At the end of each epoch, every strand in the pool is either free (colored green), cross-hybridized bonded (red), or Watson-Crick bonded (blue). During each epoch two strands are selected for potential hybridization. Here, there are three distinct interactions that may occur. The first interaction involves selecting two free strands, the second involves selecting one free and one bonded strand, and the last involves selecting two bonded strands. SynDCode decides which is most stable. If both strands are free, they will become bonded and form a duplex. If any duplex consists of a Watson-Crick

pair, it will remain bonded during the remaining epochs. If one strand is free and the other bonded, the free strand will melt the current duplex if the free energy between the free strand and one of the bonded strands is less than the free energy of the existing duplex. This procedure produces a new more stable duplex and one free strand. If both strands are bonded, both existing duplexes will melt if the free energy between one of the strands in one duplex and one of the strands in the other is less than the free energy of both current duplexes. A new more stable duplex is then formed and the other two strands become free. At each epoch at most one of these procedures may occur. The program halts when all strands reside in Watson-Crick duplexes and the code is self assembled.

5.2 Comparisons to SLSDesigner

One other well known DNA code generation software suite is SLSDesigner, produced by the Beta Lab, University of British Columbia. SLSDesigner is uses a genetic algorithm which uses the Beta Lab's Pairfold DNA thermodynamic modeler as its object function [3]. The computational complexity of Pairfold is $O(n^3)$ where n is the length of the stand(s) under consideration. SynDcode has complexity $O(n^2)$. Figure 11 shows the increased efficiency of SynDcode. Computationally more complex algorithms give more accurate pairwise computations, but *pairwise* accuracy is not necessarily the most important consideration when the primary objective is maximizing the size of a code for global thresholds. See Sections 5.2 and 5.3.

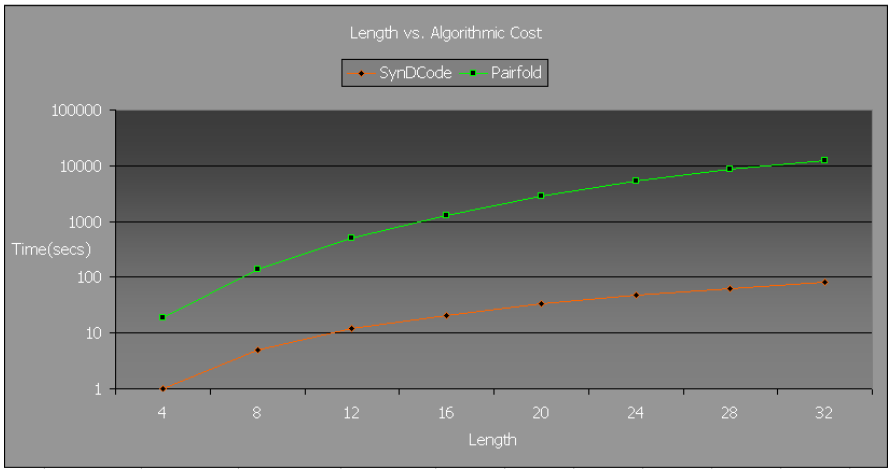

Figure 11: SynDCode and Pairfold computational time for ~1 million strand pair comparisons

5.2.1 Regression of SynDCode to Pairfold

Using SynDCode's $\psi_F(x,y), \psi^2(x,y)$, $\psi^3(x,y)$ and $\psi^4(x,y)$, a multiply linear regression was constructed with Pairfold(x,y) against X=$\psi_F(x,y)$, Y=$\psi^2(x,y)$-

$\psi^3(x,y)$ and Z=$\psi^3(x,y)$ -$\psi^4(x,y)$. The accuracy and correlation were very good. See Figures 12 and 13.

Regression Statistics	
Multiple R	0.785477
R Square	0.616974
Adjusted R Square	0.616948
Standard Percent Error	-0.05
Absolute Percent Error	0.18
Observations	44700

	Coefficients	*Standard Error*
Intercept	-4.050290292	0.023840939
Weighed 2-Stems=*X*	0.76579965	0.002948573
2'-Stem Score=*Y*	-0.825263954	0.006985618
3'-Stem Score=*Z*	-0.736910737	0.009371336

Figure 12: Pairfold vs. SynDCode(X,Y,Z)

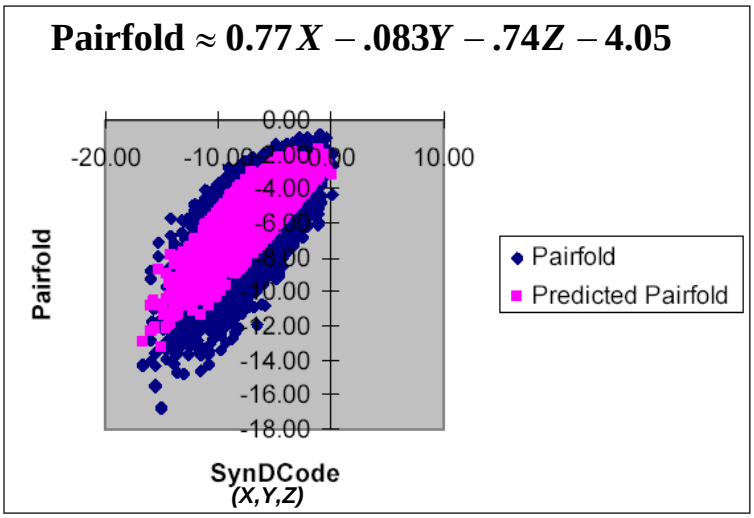

Figure 13: Pairfold vs. SynDCode multiply regression plot

5.2.2 Code Generation: SynDCode vs. SLSDesigner

Using a combination of our regression fit and greedy parsing algorithm, codes satisfying Pairfold constraints were generated using SynDCode and SLSDesigner. SynDCode's Markov generation method and increased speed performed much better than SLSDesigner's stochastic search algorithm in the more computationally complex Pairfold program. See Figure 14.

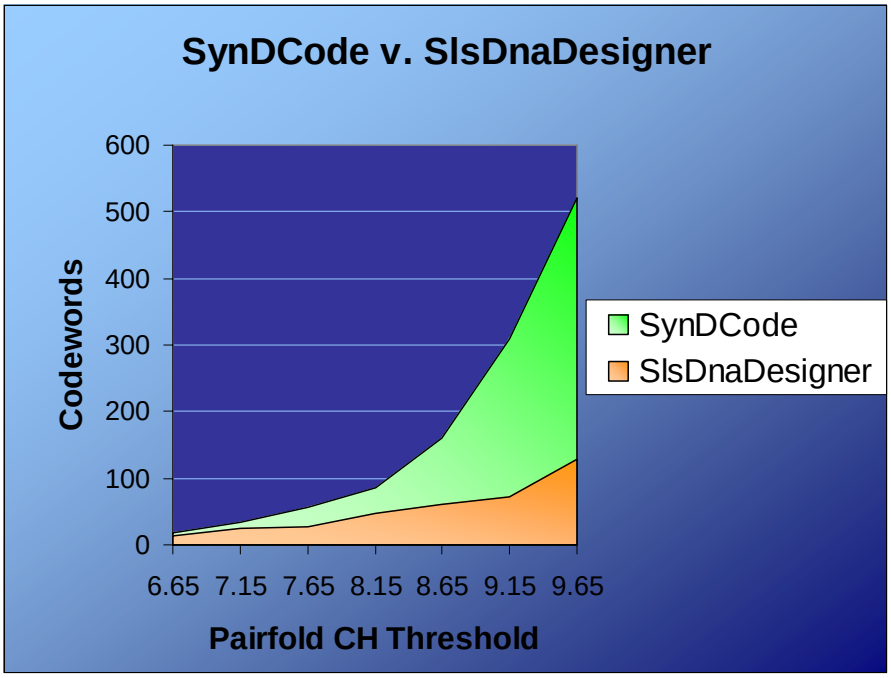

Figure 14: Comparison of DNA Code Generation Performance of SynDCode and SLSDesigner

6. References

1. A. D'yachkov, et al , A Weighted Insertion-Deletion Stacked Pair Thermodynamic Metric for DNA Codes, (with A. Dyachkov et al.), Lecture Notes in Computer Science, Springer-Verlag , Volume 3384, 90-103 (2005)

2. DNASet-Designer, available at http://ws.cs.ubc.ca/~dctulpan/dna-design.html

3. M. Andronescu, A. Condon and H. Hoos, RNAsoft, submitted to NAR for the web-based software special issue, available at http://www.rnasoft.ca/

4. M. Andronescu, Algorithms for predicting the secondary structure of pairs and combinatorial sets of nucleic acid strands, Masters Thesis, University of British Columbia, (2003).

5. A. Brenneman and A. Condon, Strand Design for biomolecular computation, Theoretical Computer Science, 287, 39-58, (2002).

6. R. Deaton, et al., A PCR Based Protocol for in Vitro Selection of Noncrosshybridizing Oligonucleotides, DNA Computing, DNA 8, M. Hagiya, A. Ohuchi (eds)., LNCS 2568, Springer, Berlin 196-204 (2002).

7. A. D'yachkov, et al., Exordium for DNA Codes, Journal of Combinatorial Optimization, 7, no.4, 369-380 (2003).

8. A. Macula, DNA-TAT Codes, USAF Technical Report, AFRL-IF-RS-TR-2003-57, http://stinet.dtic.mil/cgi-bin/fulcrum_main.pl (2003).

9. J. SantaLucia Jr., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics, Proc. Natl. Acad. Sci. USA, Vol. 95, pp. 1460-1465 (1998).

10. J. SantaLucia, Jr. and Donald Hicks, The Thermodynamics of DNA Structural Motifs, Annu. Rev. Riophys. Biomol. Struct., Vol. 33, 415-40 (2004).

11. A. Zuker, B. Mathews and C. Turner, Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide, http://www.bioinfo.rpi.edu/~zukerm/seqanal/mfold-3.0-manual.pdf (1998).